Functionality and Performance Evaluation of File Systems for Storage Area Networks (SAN)

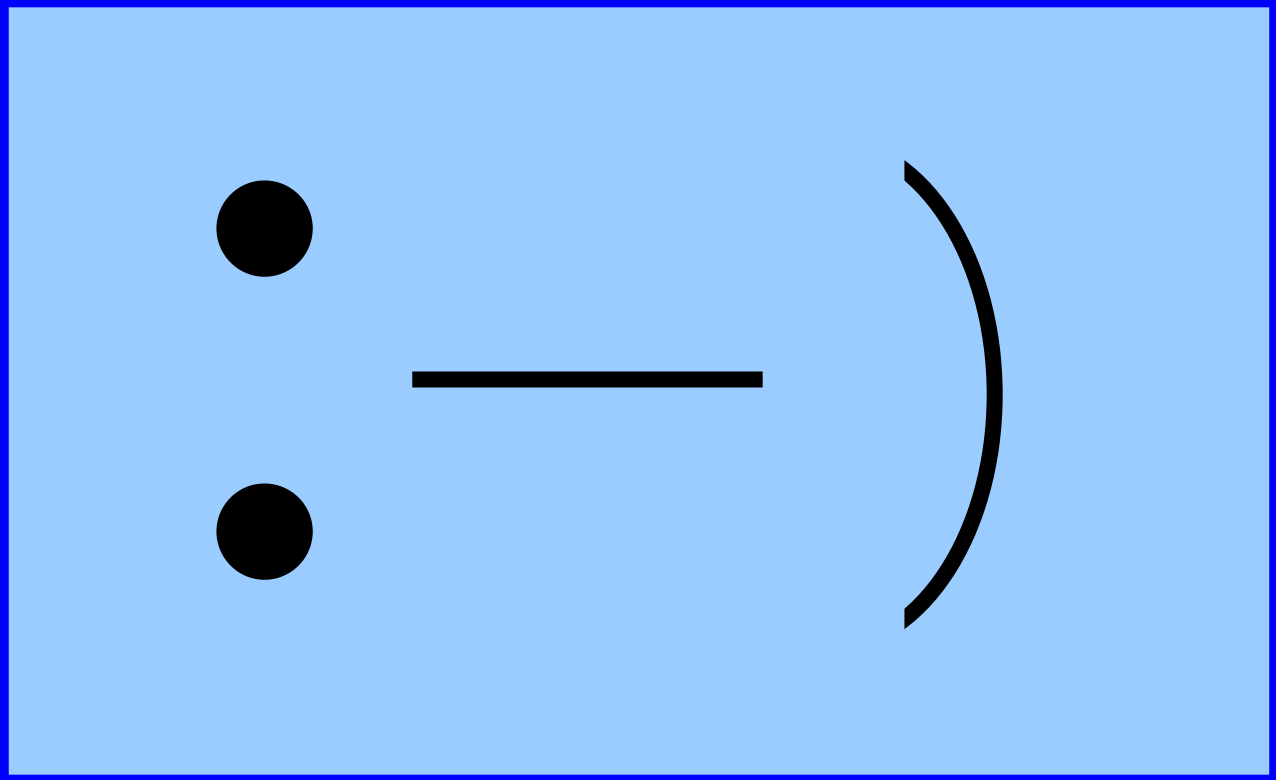
Nick Bear, Jim Finlayson, Robert Hill,
Richard Isicoff, Martha Bancroft and Hoot Thompson

Storage Technologies Knowledge Based Center,
Department of Defense

Patuxent Technology Partners, LLC

www.ptpnow.com/SANresearch

Thank You To The Committee



Agenda

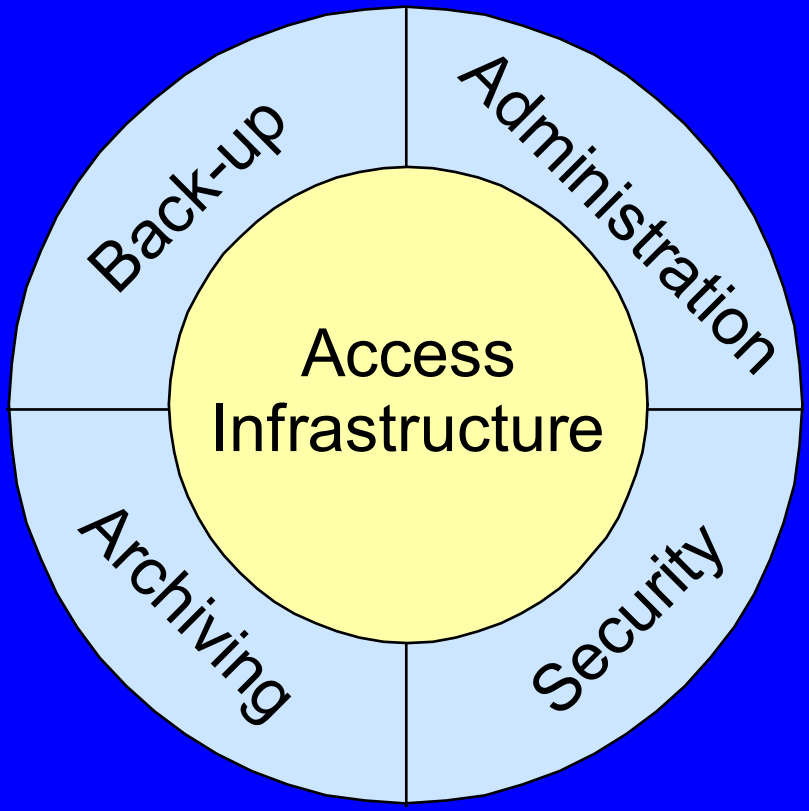
- Overview
  - Hoot Thompson
- Interim Analysis
  - Marti Bancroft
- Q&A
  - Matt O'Keefe (UMN)
  - Peter Lawthers (ADIC)
  - Steve Soltis (DataPlow)
  - Chris Stakutis (Tivoli)

SAN Birds Of A Feather

- Topics
  - Technology
  - Deployment experiences
  - Predictions
- When
  - Today
  - Immediately following last presentation
- Where
  - Main Auditorium

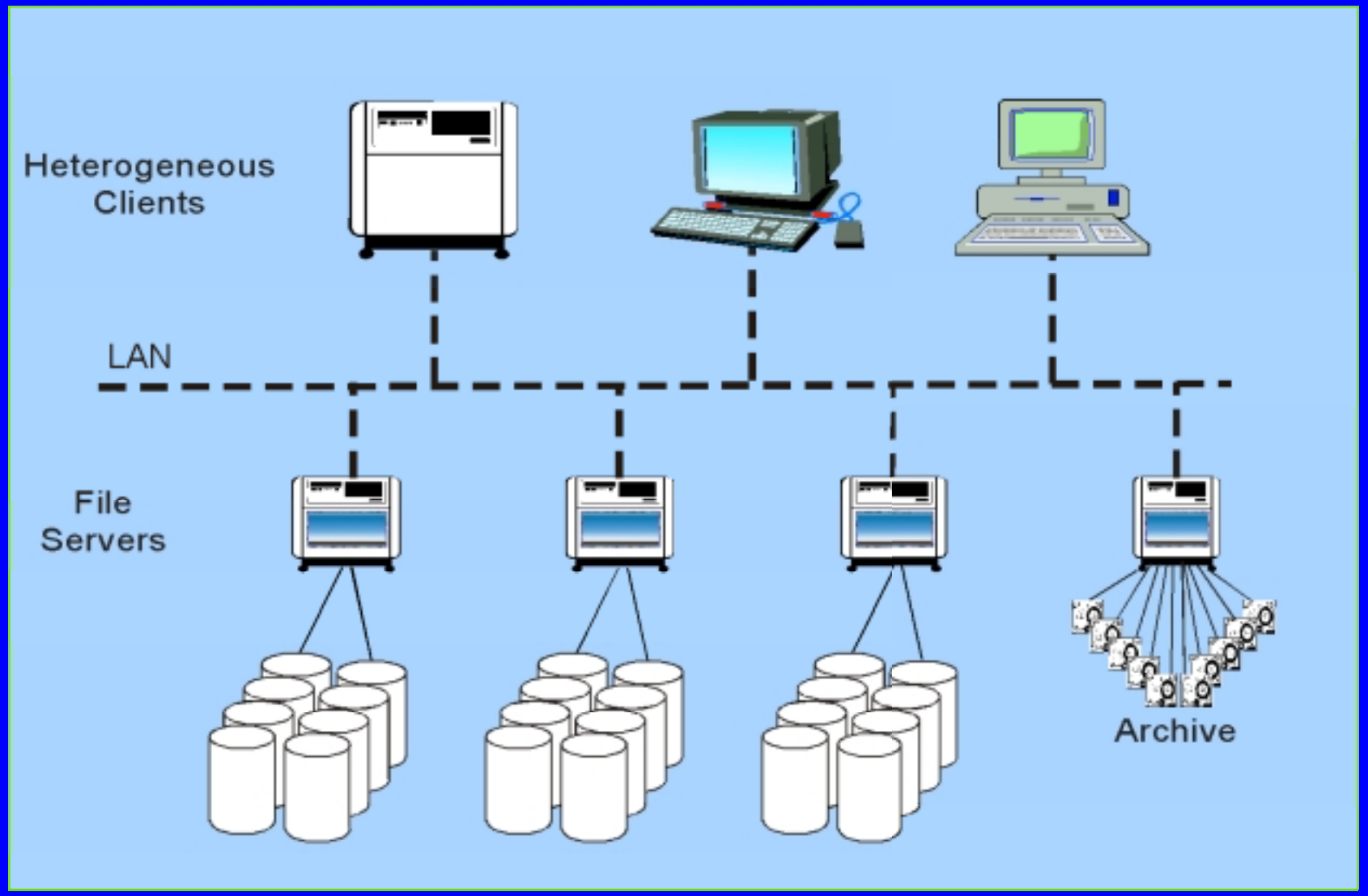
Data Management Environment

- Integrated
- Reliable
- Scalable
- Flexible

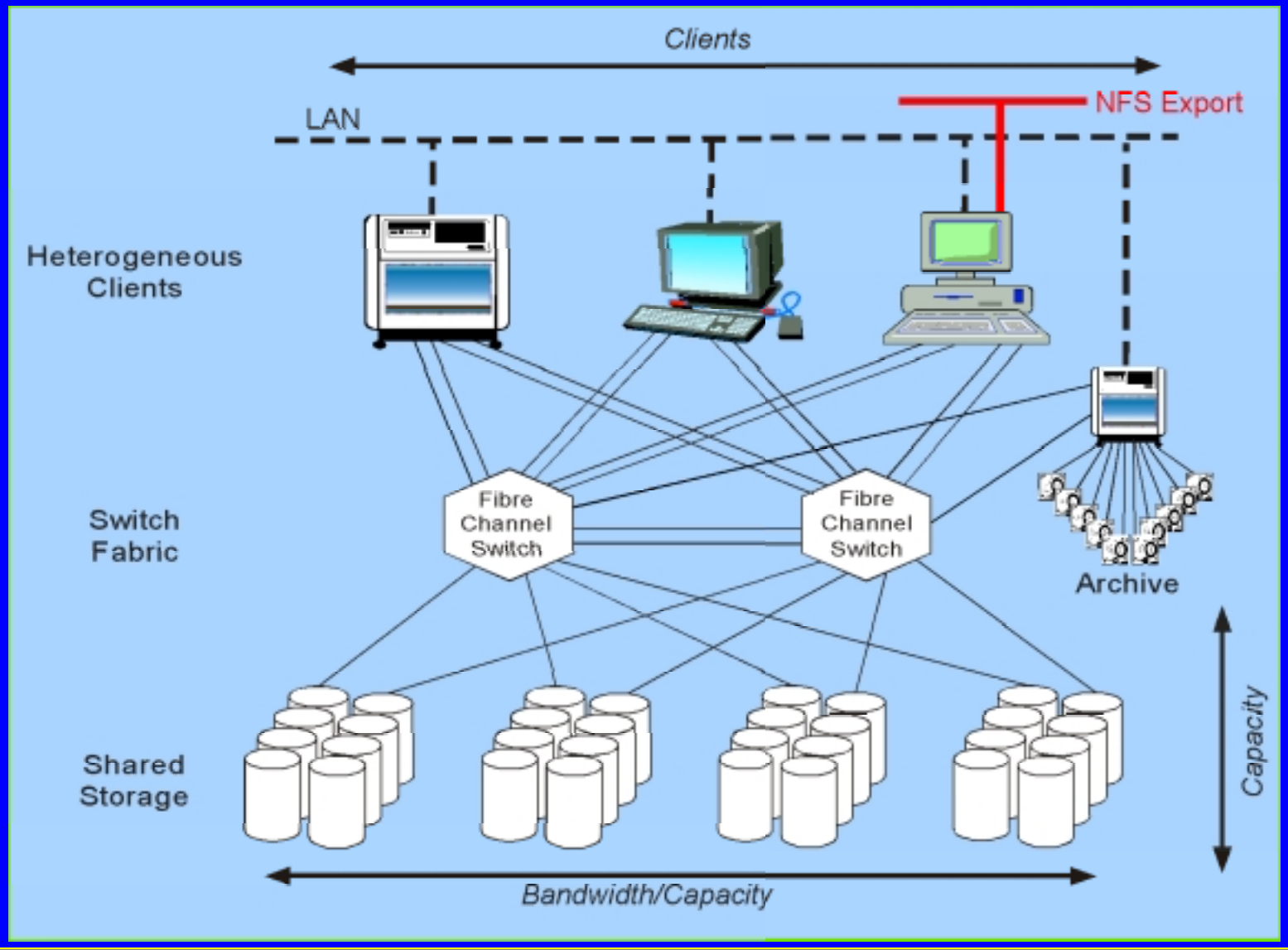


Access infrastructure is at the core

Traditional Infrastructure



SAN Based Infrastructure



Research Project Established

- Initial Objective:
  - Create an environment for evaluating SAN file systems
    - Component performance
    - Component interoperability
    - Fault tolerance, recovery and administration
  - Determine deployment roadmap
- Emerging Objective:
  - Investigate a integrated data management infrastructure
    - Back-up
    - Hierarchical storage management (HSM)
    - Centralized administration
    - Security
  - Incorporate legacy systems into new infrastructure

Sponsor Requirements

Shared concurrent reading and writing of a single file
High performance throughput for a wide range of file sizes, with an emphasis on small files
Appropriate locking mechanisms at file and sub-file level
Sustainable client bandwidth ranging from 500 megabytes/sec to 1 gigabyte/sec
High aggregate bandwidth through entire fabric (effectively equal to the number of clients times the desired per-client bandwidth)
Low latency for data access
Scaling in terms of number of clients, amount of storage, metadata management and maximum number of files supported
Transparent integration of file system into existing systems, allowing ease of use
Existing user base with support for a variety of common applications

Sponsor Requirements (cont.)

Support heterogeneous mix of operating systems
Ability to serve clients not directly attached to the SAN fabric
Additional file system functionality such as executable support, system booting, etc.
Robust SAN volume management features
HSM support
Backup support
Comprehensive set of administrative tools for configuration, monitoring and troubleshooting, allowing ease of maintainability and operation
Full range of security features
Highly available and high-integrity overall operation

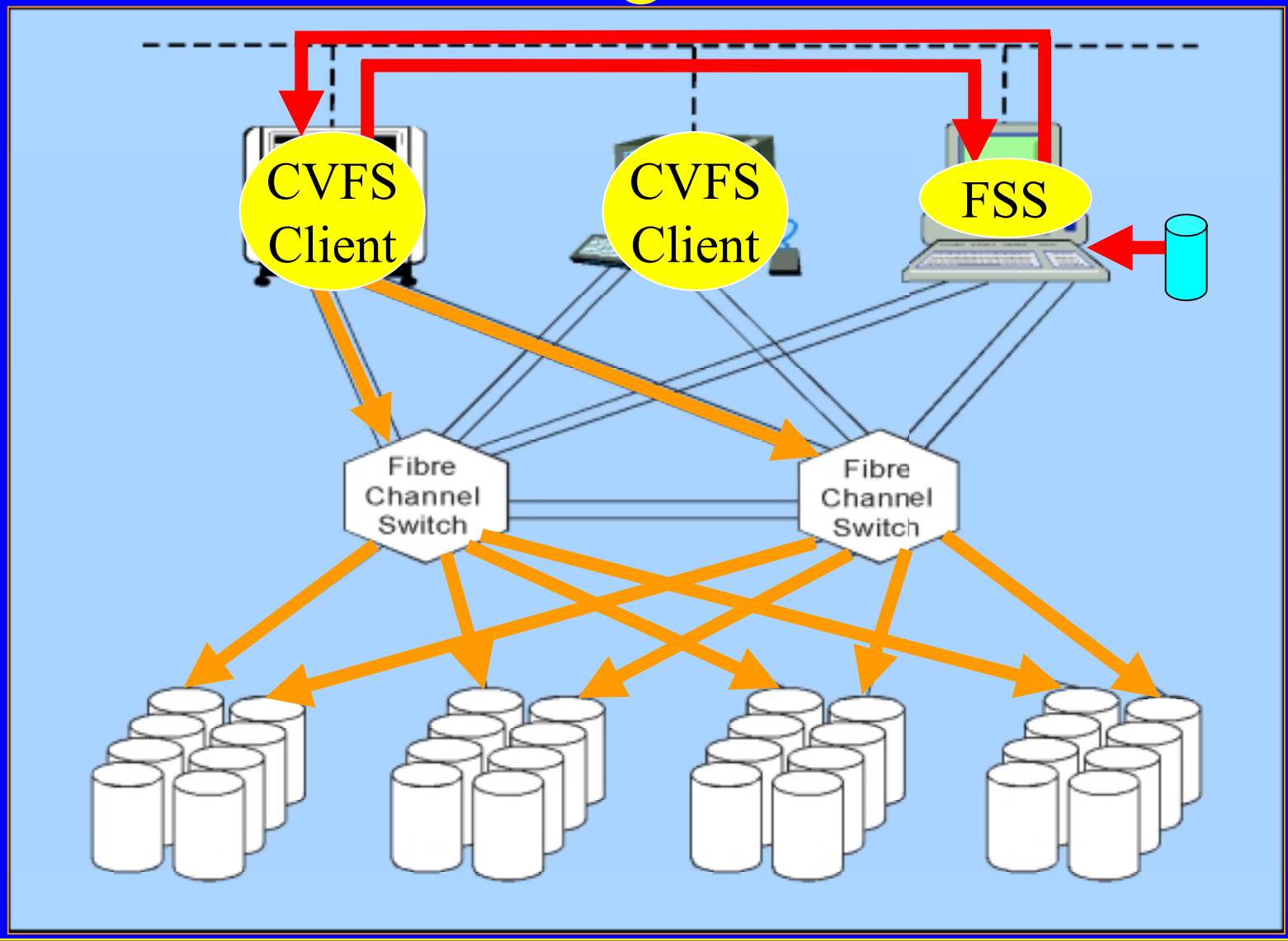
File System Summary

Product	SAN File System Design	Metadata Management	Supported Operating Systems
CVFS	Proprietary	Centralized	IRIX 6.2 to 6.5 NT 4.0 Linux (in development) Solaris (in development)
SANergy	Proprietary	Centralized	IRIX (all current releases) Solaris (all current releases) Mac 8.0+ NT 4.0 AIX (all current releases) Compaq Tru64 UNIX™ (all current releases)
DataPlow SFS	Proprietary	Centralized/ Distributed	IRIX 6.2, 6.3, 6.5 Solaris 7 and 8 Linux (future)
GFS	Open Source	Distributed	Linux

Not To Be Ignored

- Data Direct Networks, Inc.
  - Concurrent Data Networking Architecture™ (CDNA™)
- Transoft Networks, Inc.
  - FibreNet
  - Hewlett-Packard
- VERITAS® Software Corporation
- EMC Corporation

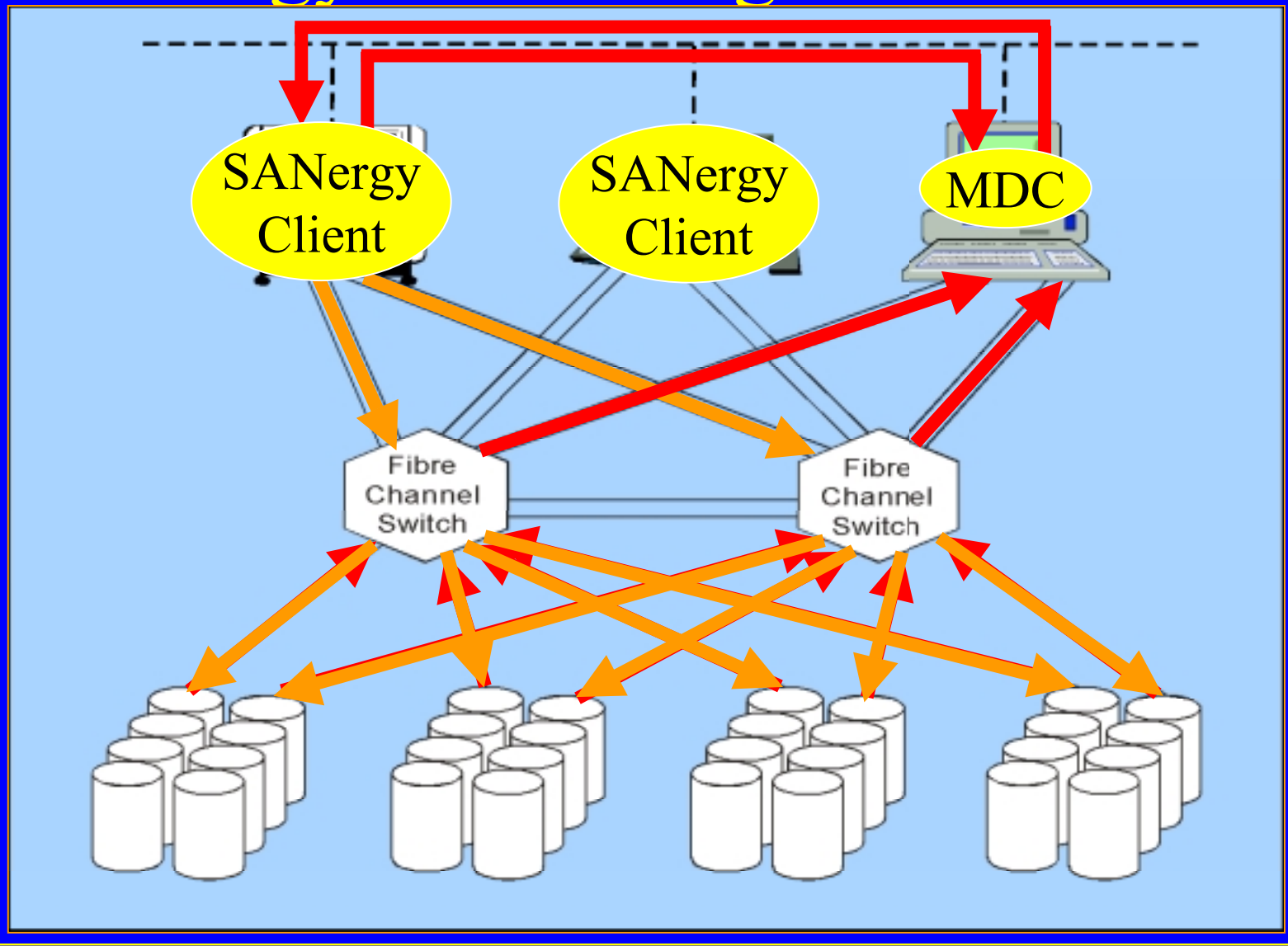
CVFS: Flow Diagram



CVFS: Features and Releases

- Key Features
  - Administrator defined striping options
    - Assignment of disks/LUNs to stripe groups
    - User defined stripe breadth
  - Special video functions
  - Guaranteed bandwidth mode for designated clients
- Upcoming releases
  - Journalled file system
  - Resiliency package for automated FSS failover
  - Integration with HSM/archive products

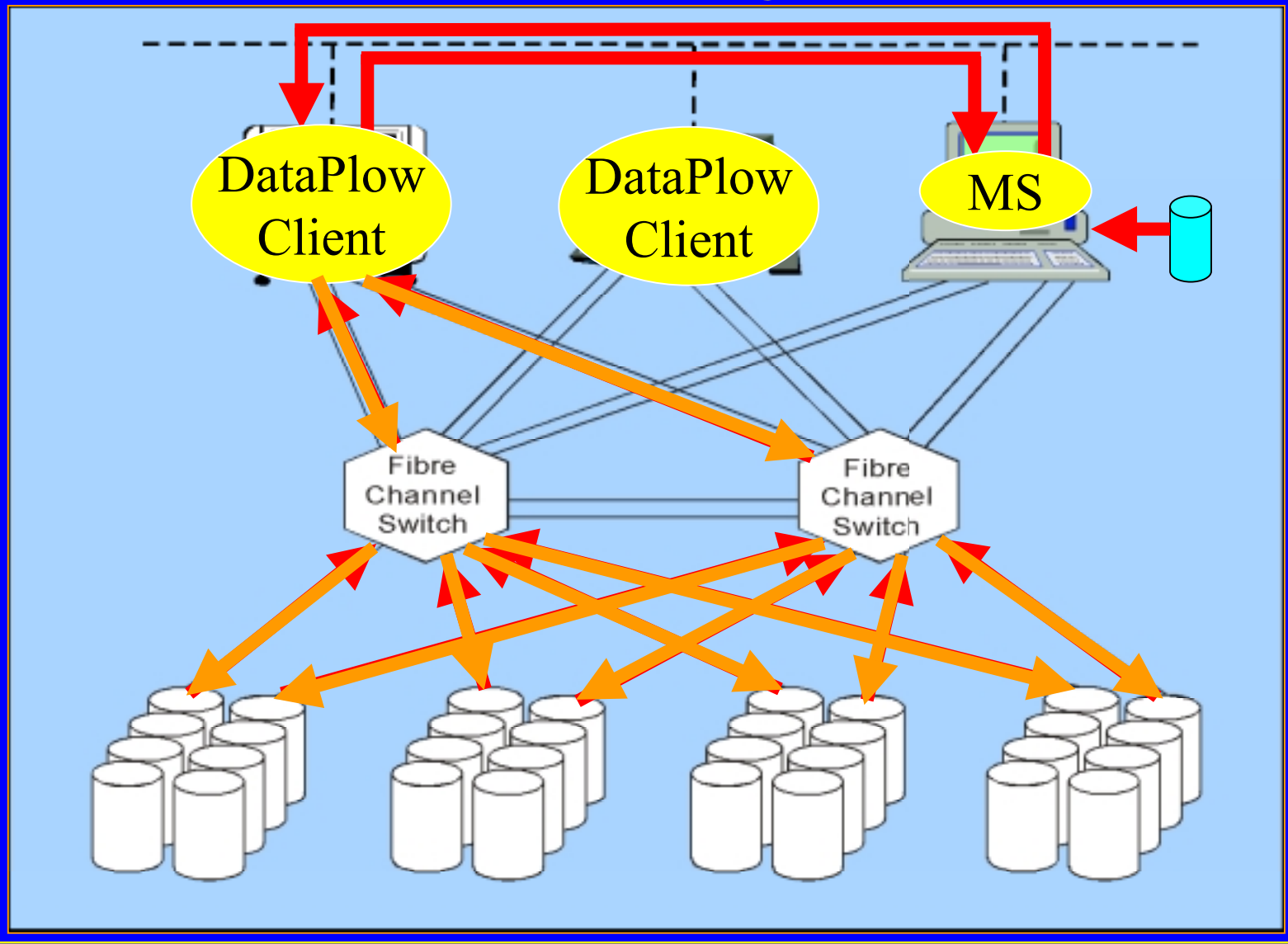
SANergy: Flow Diagram



SANergy: Features and Releases

- Key Features
  - MDC host file system provides maintenance/security
  - MDC failover option
  - Small file option
- Upcoming releases
  - Integration with LSC, Inc.'s SAM-FS
  - GUI interface for UNIX clients

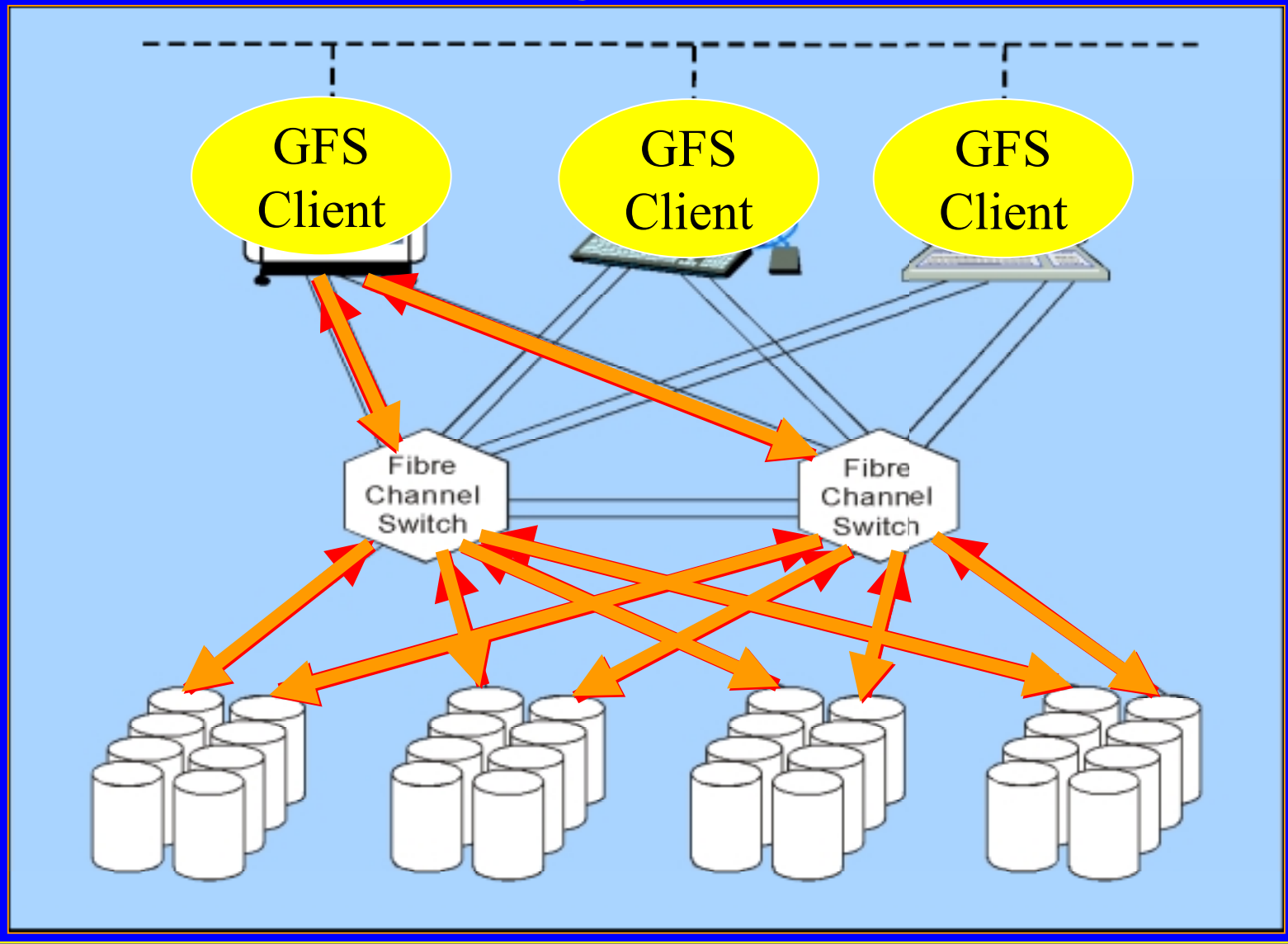
DataPlow: Flow Diagram



DataPlow: Features and Releases

- Key Features
  - File system segmentation for enhanced parallelism
  - Not every request invokes a call to the metadata server
  - Small files are stuffed into inode blocks
  - Metadata server failover using third party products
  - Journalling during allocation/dealloction
  - Two file sharing modes
    - Multiple-writers/multiple-readers
    - Single-writer/multiple-readers
- Upcoming releases
  - Heterogeneous operation

GFS: Flow Diagram



GFS: Features and Releases

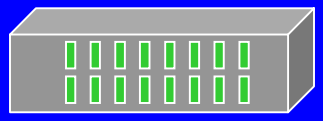
- Key Features
  - Open source model: based on Linux
    - "Core" software independent
  - Data and metadata I/O exactly like a local file system
    - Locks used to ensure only one client is modifying data at a time
  - Option for IP based global lock manager
    - Designate one of the SAN clients or standalone system
    - Compatible with third party volume management products
- Upcoming releases
  - Journalled file system
  - FreeBSD client

Observations

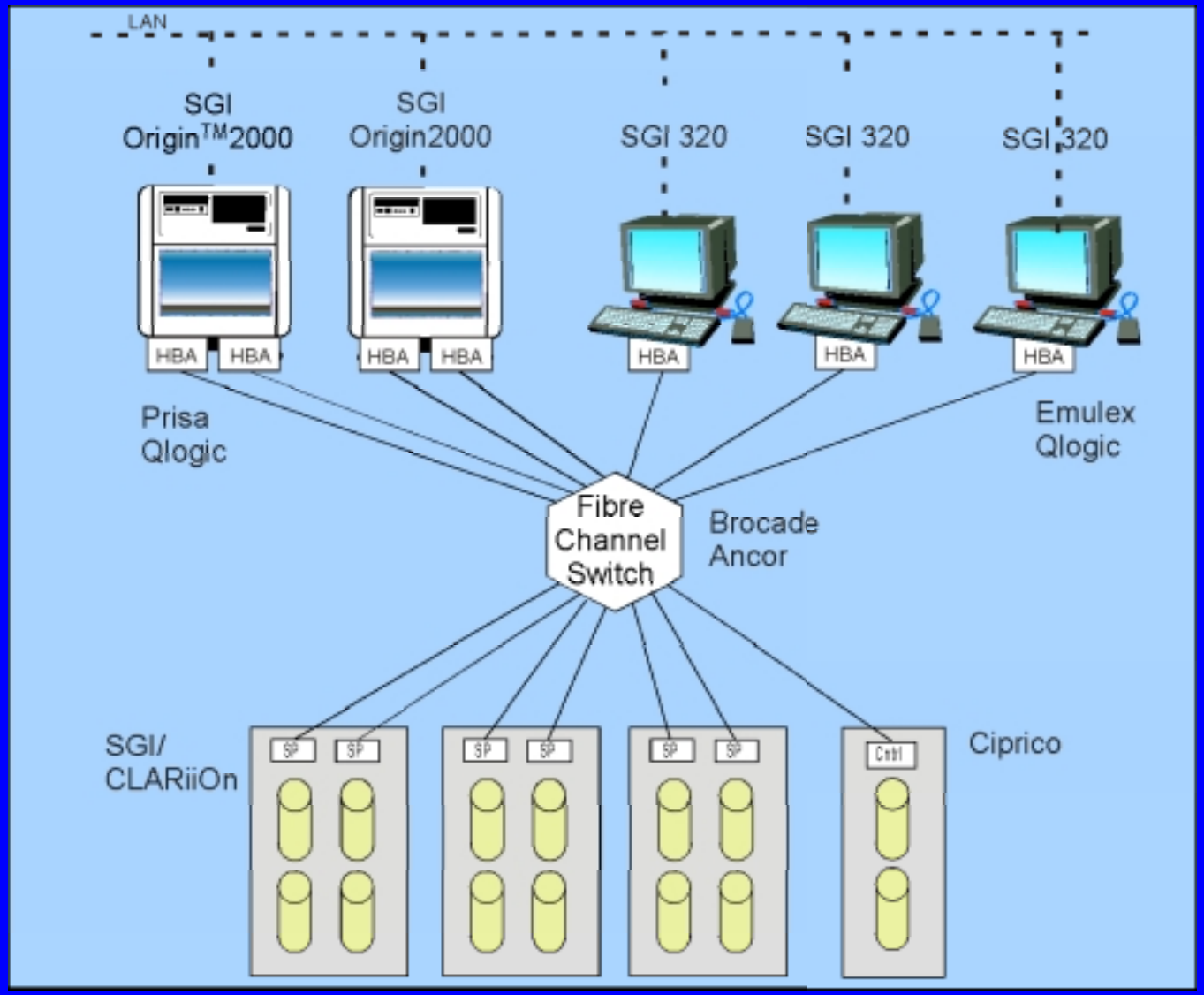
Typical Integration Experience

- RAID
  - Tuning, optimizing, etc.
  - LUN Configurations
  - MIA cycling
- Host Bus Adapter (HBA)
  - Fabric login/drivers
  - “Seeing” the RAID
  - Setting preferred paths
  - Multiple types
- Linux
  - Limited to 2.2.10 kernel
  - Numerous build issues
- Windows NT
  - Eight LUN limitation
- General
  - Reboots
  - Power cycles

Establishing and maintaining a working fabric is key!



Baseline Research Testbed



File System Testing

Product	Version	Configuration Notes
CVFS	• 1.3.8 • 1.3.10	• Heterogeneous environment • Windows NT FSS • SGI IRIX FSS
SANergy	• 1.6 • 2.0 • 2.1 (beta)	• Heterogeneous environment • Windows NT MDC • NetManage™ InterDrive® Server
DataPlow SFS	• 1.2	• Homogeneous environment • SGI Origins with console metadata server
GFS	• Antimatter Anteater	• Homogeneous environment • Global lock manager • SGI 320s with 2.2.10 kernel

Sentiments So Far……

Product	Likes	Dislikes
CVFS	• Heterogeneous • Striping options • FSS platform options • HSM/archive plans	• Limited volume management • Large file bias • Fill algorithm

SANergy	• Heterogeneous
	• Distance from the OS
	• MDC platform options
	• HSM integration plans
	• Not a file system
		• Multi-product dependency (BSOD)
		• NT MDC in a mixed environment
		  • Striping limitations
		  • No concatenation
		• Not a file system

DataPlow SFS	• Segmentation • Distributed metadata	• Homogeneous (volume management)
GFS	• Open source model • Distributed metadata • Optional lock manager	• Homogeneous • Dlock dependency (legacy systems)

All of these products are evolving!

Qualitative Assessment

- Know your workload
- No one perfect solution
  - Requirements driven
  - Asset driven
  - Budget driven
- Tuning at multiple levels
- Mileage may vary
- Simulation
  - Extend
  - SES/*workbench*®

Web Access Essential

- Product information
- Technical assistance
- Software downloads
- Community interaction

Interim Performance Evaluation

Characteristics Measured

- Data rates - isolated, combined, efficiency compared to local file system
- File manipulation rates (metadata operations)
- RAS features - behavior during problems
- Suitability for different workloads (preliminary)
- Ease of configuration, ease of use

Lessons Learned from Cray SFS

- Performance
  - Even 8000 MB/s couldn't hide some surprises
  - Separation of data and metadata accesses key to scaling
  - Tuning params very different for diverse loads
  - Not all problems due to rotating media (note that modern caching RAID controllers hide much of the disk latencies)
- Unequal Service to All...
  - Even homogeneous systems can show heterogeneous effects
  - Synchronization latency a key parameter
- RAS
  - Keep the dead from revivifying

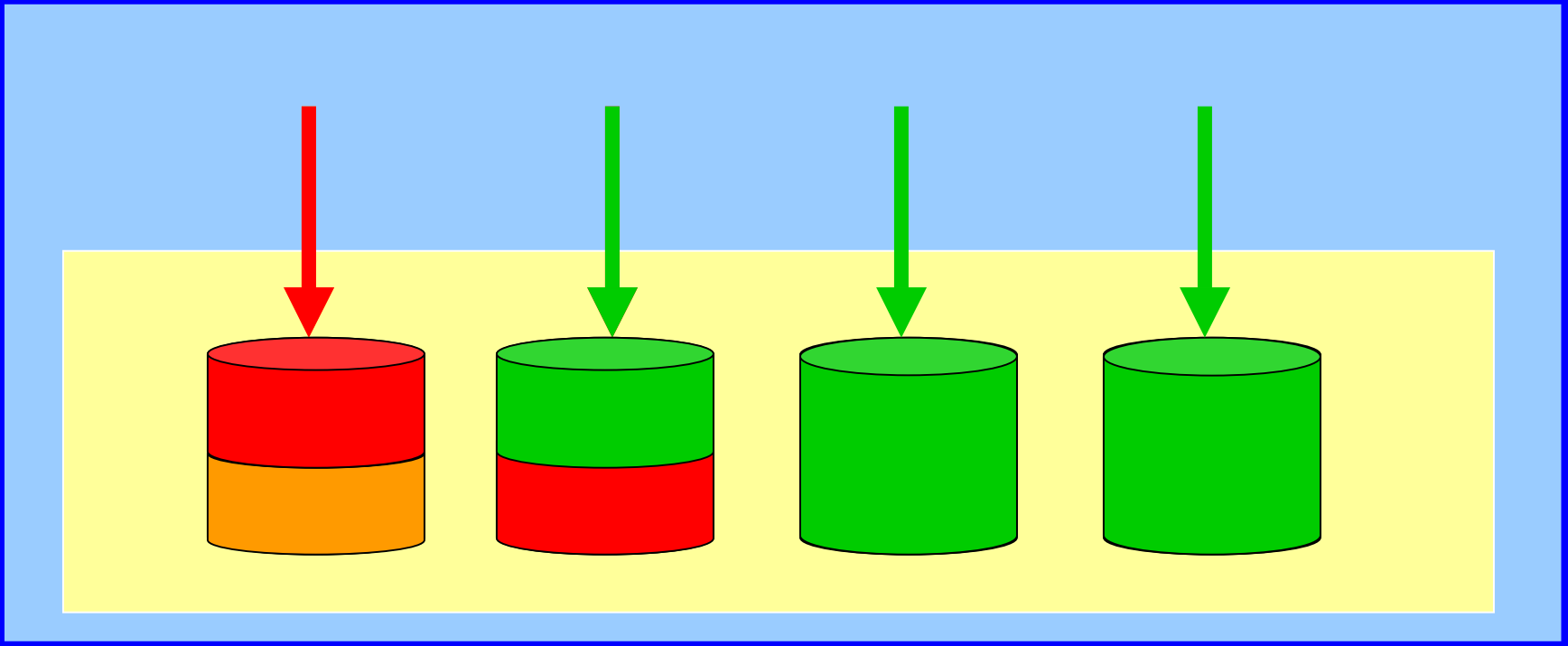
Test Programs Used

- lmdd
  - Part of lmbench
  - Available on several UNIX platforms plus NT
  - Useful for single-stream bulk data rates, NTE numbers
- xdd
  - Available on the Origins
  - Allows multiple processes, no synchronization, 1 direction only
- perf
  - SANergy only

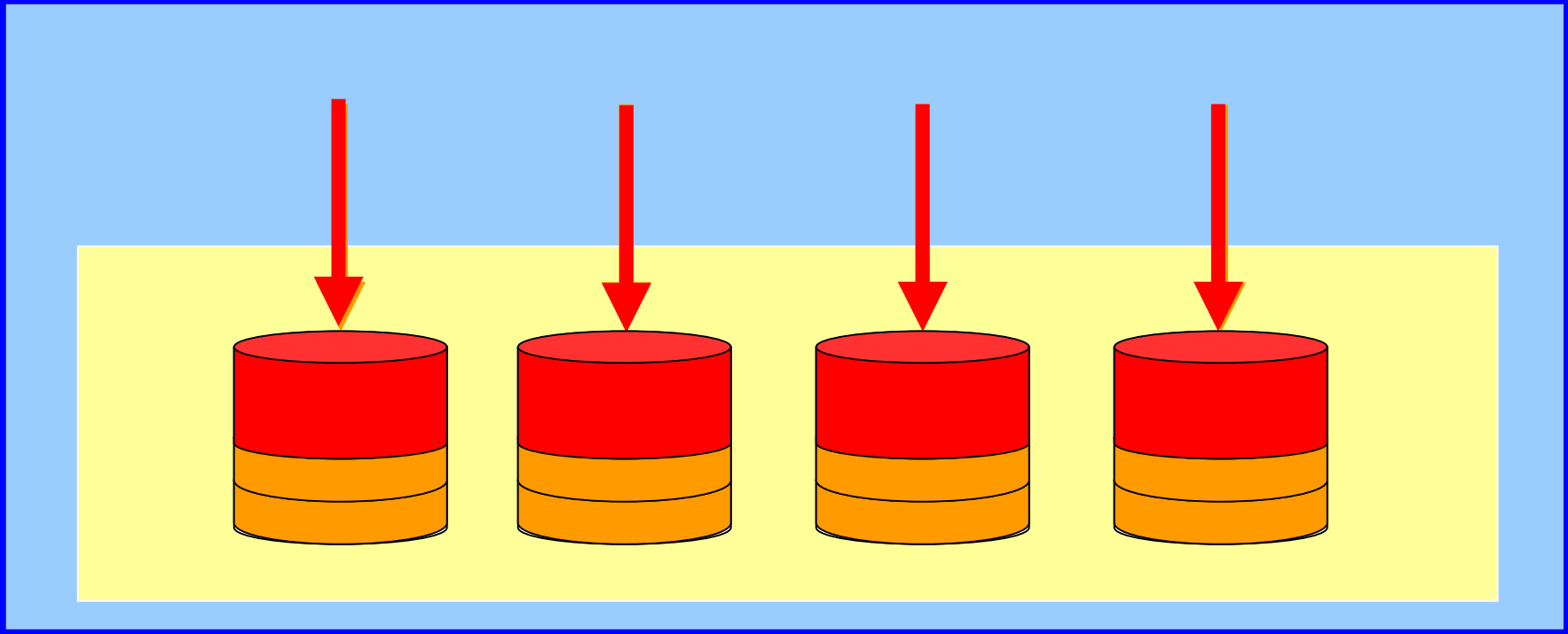
Goals of Performance Testing

- Blob performance - an efficiency measure
  - Performance comparable to local file system, raw disk
- Multiple process performance
  - Metadata handling stats for scaling models
  - File system allocation approach evaluated
  - Conflict resolution efficiency
- Tuning options vs workload type
  - Segmented concatenated appears best for small files mix
  - Memory buffering not always an option
  - Concops: 1 giant file system NOT
- RAID type vs JBOD…. Vs File System Structure

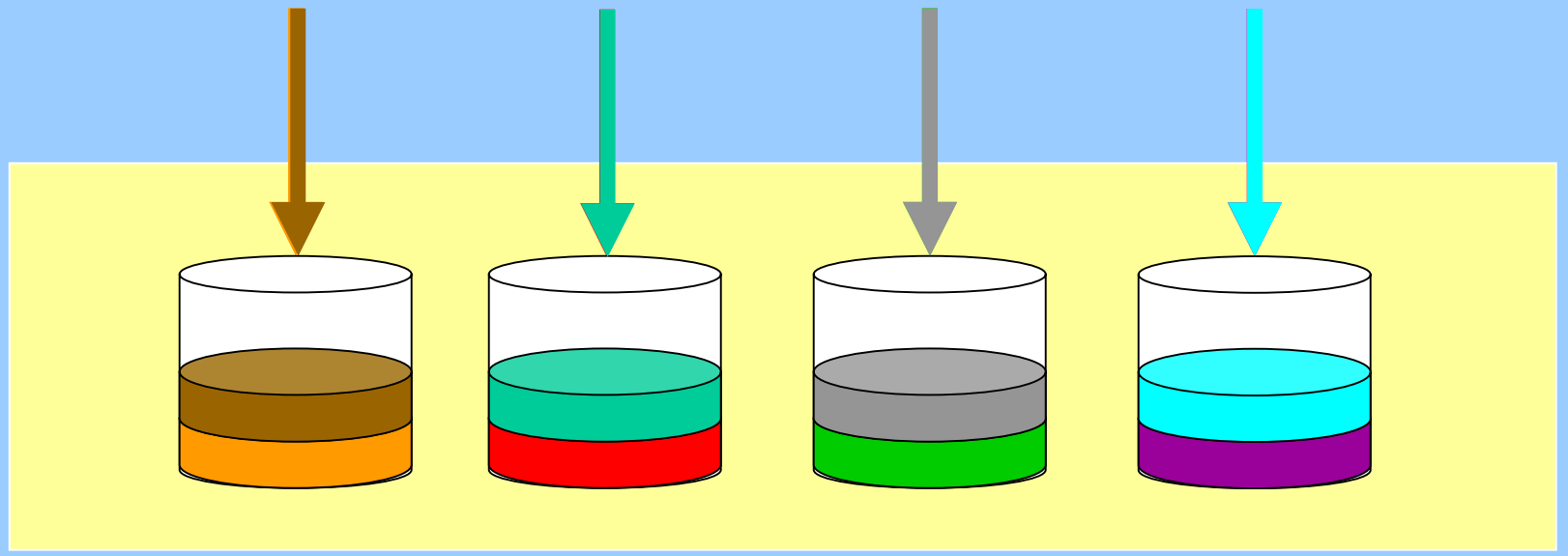
Concatenation



Striping



Segmentation



CVFS Interim Data

Cvfs - Stripe Breadth & FS Block Size (8 faster LUNs - I/O request = 32 MB)

FS Blk Size	Stripe Br	Read	Write	File Sys
4 KB	16 KB	127.6	27.18	8 x R5 9+1
4 KB	32 KB	132.54	30.08	4 SPs
4 KB	64 KB	117.98	35.39	NTFSS
4 KB	128 KB	123.51	38.23	I/O req=32MB
64 KB	32 KB	60.4	16.22	1 Origin used
TWO ORIGINS READING SIMULTANEOUSLY:				TOTAL:
4 KB	32 KB	115.14	115.39	230.53

Cvfs - Stripe Breadth & FS Block Size

(1 faster RAID-3LUN - I/O Size Varied - write cache on)

FS BS/Str Br	I/O Req	Read	Write	File Sys
4 KB/32 KB	2 MB	36.19	20.31	1 x R3 8+1
4 KB/32 KB	4 MB	36.14	20.65	1 SPs
4 KB/128 KB	2 MB	49.98	28.39	NT FSS
4 KB/128 KB	4 MB	56.59	38.76	
4 KB/128 KB	8 MB	56.66	33.11	1 Origin used
64 KB/128KB	4 MB	73.76	45.33	

Cvfs & XFS - Striped LUNs & SPs
(Prisa/Brocade/Write cache off)

#LUNs/#SPs	XFS	CVFS	% Efficiency	% sys/total	File Sys
1LUN-1SP-w	7.16	9.12	127.37	88.1/43	4 KB block
1LUN-1SP-r	30.7	33.73	109.87	" "	32 KB str br
2LUN-2SP-w	5.76	8.7	151.04	86.6/43.2	NTFSS
2LUN-2SP-r	54.78	65.37	119.33	" "	I/O req=4MB
4LUN-2SP-w	14.43	16.59	114.97	87.5/43.6	RAID-5 9+1
4LUN-2SP-r	102.15	96.67	94.64	" "	
8LUN-4SP-w	21.11	28.93	137.044055	74.5/43.0	
8LUN-4SP-r	62.57	134	214.160141	" "	

Cvfs - New Version - Stripe Breadth & LUN Type

(Qlogix/Brocade/write cache on/I/O req = 4MB)

Stripe Br	RAID-5 8+1	RAID-5 8+1	RAID-3 8+1	RAID-3 8+1	File Sys
32	16.13 - w	21.79 - r	28.21 - w	53.48 - r	4KB Block
128	16.75 - w	27.36 - r	52.58 - w	68.83 - r	I/O req = 4MB
256	25.57 - w	34.15 - r	75.30 - w	80.16 - r	same file
512	36.27 - w	43.18 - r	76.28 - w	87.30 - r	
1024	41.38 - w	52.37 - r	77.07 - w	86.84 - r	
2048	41.53 - w	52.20 - r	77.02 - w	87.08 - r	

Cvfs - New Version - lmdd - Striped LUNs

(Qlogix/Brocade/write cache on/I/O req = 4MB)

#LUNs/#SPs	LUN type	Write	Read	File Sys
4 LUNs/4 SPs	RAID-3 8+1	89.01	94	4KB Block
2 LUNs/1 SP	RAID-5 4+1	31.36	40.33	1024 str br
1 LUN/1 SP	RAID-5 8+1	41.37	52.35	same file
				I/O req=4MB

Cvfs - New Version - xdd

(Qlogix/Brocade/write cache on)

Block Size	Low Rate	High Rate	Aggregate	LUN Type
1 MB	4.46	4.85	142.62	4 x R3 8+1
2 MB	4.4	4.9	140.88	" "
3 MB	4.48	5.12	143.28	" "
4 MB	4.54	5.11	145.29	" "
1 MB	0.72	1.05	22.89	1 XR5 8+1

SANergy Interim Data

SANergy Blob Efficiency - 1 proc
(SANergy 1.6/Prisa/Brocade)

I/O req size	Host	Rate	LUN type	Test Prog
1 MB	menagea	25.0 - wr	1 x R3 8+1	perf (1 proc)
1 MB	menagea	60.2 - rd	"	perf
2 MB	menagea	35.3 - wr	"	perf
2 MB	menagea	56.9 - rd	"	perf
4 MB	menagea	41.0 - wr	"	perf
4 MB	menagea	48.8 - rd	"	perf
4 MB	menagea	42.7 - wr	"	perf
4 MB	menagea	51.2 - rd	"	perf
6 MB	menagea	39.4 - wr	"	perf

SANergy Blob Efficiency vs XFS
(SANergy 1.6/Prisa/Brocade)

I/O req size	XFS	SANergy	% efficiency	Config.
2MB write	36.25	35.3	97.38	perf
2 MB read	77.39	56.9	73.52	each LUN =
4 MB write	45.44	42.7	93.97	RAID-3 8+1
4 MB read	72.96	51.2	70.18	XFS defaults
				SANergy also

SANergy Blob Efficiency - 4 proc
(SANergy 1.6/Prisa/Brocade)

men_a_0_0	men_a_0_1	men_a_1_0	men_a_1_1	Aggregate	Config.
36.6 - wr	37.9 - wr	36.6 - wr	37.9 - wr	149.00 MB/s	perf
37.9 - rd	36.6 - rd	36.6 - rd	35.3 - rd	146.4 MB/S	each LUN =
37.9 - WR	37.9 - RD	37.9 - WR	39.9 - RD	153.6 MB/s	RAID-3 8+1

SANergy Blobs - lotsa procs
(SANergy 1.6/Prisa/Brocade)

menagea	menageb	Aggregate	Config:
27.7	27.7	all writing	perf
27.7	27.7		O's to 8 LUNs
27.7	28.4		4 MB I/O req
28.4	28.4	223.7	R3 8+1 LUNs
			1 FS/LUN

SANergy - Striped - 8 x RAID-3 8+1
(SANergy 1.6/Prisa/Brocade)

menagea	menageb	NT - I/O size	Aggregate	Config:
47.32 - wr	47.33 - wr	37.29 -wr-8MB	131.94 MB/s	perf
47.61 - rd	48.91 - rd	66.48 -rd-32MB	163.00 MB/s	striped 8 LUNs
-	-	52.83-rd-8MB		4 MB I/O req
75.0 - rd	-	-		R3 8+1 LUNs
-	-	73.86-rd-32MB		1 FS across all
-	75.2 - rd			

SANergy - Latest Version - 1 LUN

I/O req size	Host	Rate	LUN type	Test Prog
1 MB	menagea	58.09 - wr	1 x R3 8+1	perf
1 MB	menagea	60.80 - rd	"	perf
4 MB	menagea	56.38 - wr	"	perf
4 MB	menagea	52.68 - rd	"	perf

SANergy - Latest Version - Cache Benefit

- 128KB writes - 23 MB/s to 330 MB/s once in memory cache (1 MB file)
- 128 KB reads - 22 to 29 MB/s (1 MB file) - cache benefit unclear though configured for read and write
- 1 GB file to compare: 22.83 MB/s write, 27.36 MB/s read

DataPlow Interim Data

DataPlow - Blob Efficiency vs Raw

lmdd - striped 4 x RAID-3 8+1 LUNs

I/O req size	xlv raw	16 seg FS	% Efficiency
1 MB write	67.29	66.19	98.37
1 MB read	126.59	127.6	100.80
2 MB write	67.2	67.16	99.94
2 MB read	127.78	128.3	100.41
3 MB write	66.92	67.48	100.84
3 MB read	138.29	138.54	100.18
4 MB write	67.55	67.49	99.91
4 MB read	145.56	145.7	100.10

DataPlow - Segmentation & Stripes

xdd - 4 x RAID-3 8+1 LUNs

I/O Req Size	# procs	16 segments	1 segment
1 MB	1 writer	66.19 MB/s	64.89 MB/s
"	2 writers	35.35 MB/s	54.70 MB/s
"	4 writers	38.87 MB/s	55.87 MB/s
"	1 reader	127.60 MB/s	126.39 MB/s
"	2 readers	63.03 MB/s	75.05 MB/s
"	4 readers	56.02 MB/s	72.02 MB/s

DataPlow - Segmentation & Stripes

xdd - 4 x RAID-3 8+1 LUNs

I/O Req Size	# procs	16 segments	1 segment
4 MB	1 writer	67.49 MB/s	67.22 MB/s
"	2 writers	60.80 MB/s	65.88 MB/s
"	4 writers	60.14 MB/s	64.74 MB/s
"	1 reader	145.70 MB/s	145.51 MB/s
"	2 readers	104.22 MB/s	117.74 MB/s
"	4 readers	124.51 MB/s	110.04 MB/s

DataPlow - Blob Efficiency- 1 LUN

(Compare to other 1 x RAID-3 8+1 LUN data)

menagea	menageb	BS/FS(lmdd)	Aggregate
-	70.79 - wr	4MB/ 1GB	70.79 MB/s
71.60 - wr	-	" "	71.60 MB/s
87.14 - rd	-	" "	87.14 MB/s
-	87.26 - rd	" "	87.26 MB/s
35.43 - wr	35.51 -wr	4MB/ 2GB	70.94 MB/s
34.74 - rd	36.03 - rd	" "	70.77 MB/s
35.43 - rd	44.86 - wr	" "	80.20 MB/s

DataPlow - MultiProcess/MultiHost
(xdd on default file systems, 4 LUNs)

FS Type	32 readers	1 writer	xdd total	32 writers
Striped -SWr	50.71	32.43	83.14	53.59
Striped -Mwr	36.08	5.18	41.26	42.26
Concatenated	113.54	8.94	122.48	185.11
1 MB I/O				4 MB I/O

Future Testing

- Subset of SAN products, plus others (tbr)
- Increased Number of SAN clients, more OS's
- More metadata details for scaling models
- More tuning studies for mixed workloads
- Add HSM to mixture
- If necessary, test program development to better reflect customer's requirements (haven't found the 'perfect test suite' yet!)

Informational URLs

- Storage Network Industry Association
  - www.snia.org
- Fibre Channel Industry Association
  - www.fibrechannel.org
- CentraVision File System
  - www.centravision.com
- SANergy
  - www.sanergy.com
- DataPlow, Inc.
  - www.dataplow.com
- Global File System
  - www.globalfilesystem.org

Cvfs - Stripe Breadth & FS Block Size (1 faster LUN - I/O request = 4 MB)

FS Blk Size	Stripe Br	Read	Write	File Sys
4 KB	32 KB	35.23	11.87	1 x R5 9+1
				1 SPs
				NTFSS
				I/O req=4MB

Cvfs & XFS - Single RAID-3 8+1 LUN (Prisa/Brocade/Write cache on)

I/O req size	XFS	CVFS	% Efficiency	File Sys
2 MB - w	36.25	28.39	78.32	4 KB block
2 MB - r	77.39	49.98	64.58	128 KB str br
4 MB - w	45.44	38.76	85.30	NTFSS
4 MB - r	72.96	56.59	77.56	RAID-3 8+1

Cvfs - Wider Stripe Multi-Client
(Prisa/Brocade/Origin FSM/ same file)

Stream 1	Stream 2	Aggregate	Read Blk Size	File Sys
97.02		97.02	4 MB	4 x R3 4+1
96.87		96.87	8 MB	4 SPs
58.83	58.31	117.14	4 MB	BS=4kB
58.85	58.83	117.68	4 MB	StBr=64KB
55.91	55.93	111.84	8 MB	O2K FSS
53.58	53.58	107.16	2 MB	
58.98	58.93	117.91	4 MB	

Cvfs - Multiple Client Effects

4MB/Prisa/Ancor/write cache off

Stream 1	Stream 2	Aggregate	Same File	FileSys
41.59	42.3	83.89	r+r	2 xR3 4+1
5.57	5.4	10.97	w+w	2 SPs
21.68	5.3	26.98	r+w	BS=4kB
2 File Systems, 2 FSSs (one NT, one Origin):				StBr=16KB
51.75	38.71	90.46	r+r	NT FSS

Cvfs - Efficiency vs XFS
(Prisa/Ancor/write cache on, 1 - RAID-3 4+1)

I/O Block	XFS	CVFS	% Efficiency	Direction
1 MB	13.19	21.24	161.03	write
1 MB	26.56	26.72	100.60	read
2 MB	17.22	21.92	127.29	write
2 MB	28.34	27.82	98.17	read
4 MB	28.84	30.15	104.54	read

SANergy: Features Tested

- Centralized metadata control: Metadata Controller (MDC)
  - Hosted on a standalone system or combined with SAN client
  - Tested on NT and Origin (IRIX) using NTFS
  - Metadata mingled/stored on SAN with file data
  - Requires fabric connectivity for NT system
- Partly split data versus control flow
  - Control data passed between SAN clients and MDC via LAN
  - Data files redirected to fibre channel links by ‘fusing’
  - Small by measurable metadata transactions during writes
- MDC volume management: disks/LUNs labeled as NTFS entities

SANergy: Tuning Options

- Key Parameters
  - Striping (NT parameters only)
  - In-memory buffering on clients (Origin) means small files may see memory-memory transfer rates
  - Fusing (either ‘on’ or ‘off’; off = NFS rates)
  - Small file mode (limited testing)
- Opportunities
  - Concatenated file systems?
  - Improved client performance

SANergy Blob Efficiency - NT
(SANergy 1.6/Prisa/Brocade)

I/O req size	Host	Rate	LUN type	Test Prog
1 MB	NT	16.35 - wr	1 x R3 8+1	lmdd
1 MB	NT	50.03 - rd	"	"
2 MB	NT	21.11 - wr	"	"
2 MB	NT	47.20 - rd	"	"
32 MB	NT	71 - rd	"	perf

SANergy Blob Efficiency - 2 proc
(SANergy 1.6/Prisa/Brocade)

men_a_0_0	men_a_0_1	Aggregate	Config:
41.0 - wr	42.7 - wr	83.7 MB/s	4 MB I/O req
48.8 - rd	48.8 - rd	97.6 MB/s	perf
42.7 - wr	48.8 - rd	91.5 MB/s	each LUN =
			RAID-3 8+1

SANergy Blob Efficiency - 2 O2Ks
(SANergy 1.6/Prisa/Brocade)

menagea	menageb	Aggregate	Config:
27.7 - wr	26.9 - wr	54.6 MB/s	perf
32.0 - rd	32.0 - rd	64.0 MB/s	same LUN
42.7 - wr		42.7 MB/s	diff. Files
4 MB I/O req			R3 8+1

SANergy Blob Efficiency - 3 clients (SANergy 1.6/Prisa/Brocade)

menagea	menageb	NT (LUN K)	Aggregate	Config:
-	42.7 - wr	-	42.7 MB/s	perf
42.7 - wr	21.8 - wr	33 - wr	64.5/ 33	O's same LUN
44.5 - wr	21.3 - wr	33.04 - wr	65.8/ 33.04	2 Files,4MB
-	42.7 - wr	-	42.7 MB/s	R3 8+1
42.7 - wr	42.7 - wr	-	85.4 MB/s	NT=own LUN
44.5 - wr	-	37.5 - wr	44.5/ 37.5	NT=16 MB
-	21.3 - wr	36.62 - wr	21.3/ 36.62	I/O reqs
-	23.3 - wr	21.77 - wr	23.3/ 21.77	NT=4MB here

SANergy Blobs - lotsa procs
(SANergy 1.6/Prisa/Brocade)

menagea	menageb	Aggregate	Config:
30.1	30.1	all reading	perf
30.1	29.3		O's to 8 LUNs
30.1	30.1		4 MB I/O req
29.3	29.3	238.4	R3 8+1 LUNs
			1 FS/LUN

DataPlow: Features of Interest

- Distributed (split) metadata
  - Higher level namespace-type information managed by metadata server
  - Extent-level data stored directly on the shared storage
  - SAN Clients allocate/deallocate their own metadata and real data
  - Small files contained within the metadata block on shared disk
- Metadata server
  - Hosted standalone or combined with SAN client
  - Runs on IRIX or Sun platform
  - Part of metadata stored on host's local file system disk
- Split data versus control flow
- Segmentation, striping vs concatenated file systems

DataPlow: Tuning Parameters

- Key Variables
  - File system block size, cache sizes, allocation parameters
  - Segments to distribute traffic
  - Multi writer/readers vs single writer/multireader
  - Striped (stripe breadth controlled by XLV) vs concatenated
- Encouraging obervations
  - Corrupted file system continues correct operation for data areas not damaged
  - Segmentation improves work distribution on concatenated file systems dramatically

DataPlow - Segmentation & Stripes

xdd - 4 x RAID-3 8+1 LUNs

I/O Req Size	# procs	16 segments	1 segment
2 MB	1 writer	67.16 MB/s	65.66 MB/s
"	2 writers	53.96 MB/s	63.58 MB/s
"	4 writers	53.32 MB/s	62.22 MB/s
"	1 reader	128.30 MB/s	128.07 MB/s
"	2 readers	81.59 MB/s	95.61 MB/s
"	4 readers	89.72 MB/s	93.62 MB/s

DataPlow - Segmentation & Stripes

xdd - 4 x RAID-3 8+1 LUNs

I/O Req Size	# procs	16 segments	1 segment
3 MB	1 writer	67.48 MB/s	66.96 MB/s
"	2 writers	58.99 MB/s	65.18 MB/s
"	4 writers	57.89 MB/s	63.70 MB/s
"	1 reader	138.54 MB/s	138.70 MB/s
"	2 readers	104.22 MB/s	106.45 MB/s
"	4 readers	124.51 MB/s	101.99 MB/s

DataPlow - xlv stripe breadth check

xlv stripe	menagea	BS/File Size
64x512	128.24 - rd	4 MB/ 2 GB
128 x 512	146.62 - rd	" "
256x512	136.72 - rd	" "
512x512	139.65 - rd	" "
1024x512	129.79 - rd	" "
2048x512	82.64 - rd	" "

DataPlow - Striped - Single/MultiWriter (lmdd - 4 x RAID-3 8+1 LUNs - 128x512)

FS Type	Menagea	Menageb	Aggregate
S-Writer	66.04 - wr	-	66.04 MB/s
M-Writer	17.93 - wr	-	17.93 MB/s
S-Writer	69.90 - rd	69.80 - rd	139.70 MB/s
M-Writer	36.21 -rd	36.08 - rd	72.29 MB/s
S-Writer	33.34 - wr	33.28 - wr	66.62 MB/s
M-Writer	23.62 - rd	21.46 - wr	45.08 MB/s

Other URLs Of Interest

- Benchmarks
  - General (Postmark, BONNIE, etc.)
    - devlinux.com/projects/reiserfs/bens.html
  - lmdd
    - www.bitmover.com/lm_engr/lmbench/
- Simulation
  - SES/*workbench*®
    - www.ses.com
  - Extend
    - www.imagethatinc.com

Recommended Reading

- Building Storage Networks
  - By Marc Farley

- Designing Storage Area Networks
  - By Tom Clark

Six Degrees of SAN

- Variable number of connections per client
  - Driven by required client bandwidth
  - Allows connection striping
- Flexible switch fabric
  - Increase client and storage connectivity
  - Increase system bandwidth
  - Improve availability
  - Connect both disk and tape
- Expandable shared storage
  - Increase capacity by adding more spindles
  - Increase bandwidth by adding more controllers

SAN State of the Union

- Acknowledged potential
  - Performance comparable to directly attached storage
- Technology is maturing rapidly
  - Shared file systems
  - Fibre channel-based interconnect solutions
  - Management and administration tools
- Experiencing broad industry participation
  - Computer manufacturers
  - Connectivity providers
  - System integrators
- Corporate consolidation

When will the technology be ready for full scale deployment?

CVFS: Basic Architecture

- Centralized metadata management: File System Services
  - Standalone host or combined with a SAN client
  - Windows NT® or SGI IRIX™ platform
  - Metadata stored on host's local file system
- Split data versus control flow
  - Control data passed between SAN clients and FSS via LAN
  - Data files moved over fibre channel links
- Volume management: disks/LUNs labeled as CVFS entities
  - FSS builds volumes if connected to fabric
  - Data stored in CVFS file format
  - SAN client builds volumes if FFS hosted on standalone server
- 64-bit file system
- Normal file system utilities such as cvfsck

SANergy: Basic Architecture

- Centralized metadata control: Metadata Controller (MDC)
  - Standalone host or combined with SAN client
  - NTFS, Solaris UFS or Quick File System (QFS) platform
  - Metadata mingled/stored on SAN with file data
  - SAN connectivity required
- Split data versus control flow
  - Control data passed between SAN clients and MDC via LAN
  - Data files redirected to fibre channel links
- MDC volume management:
  - Disks/LUNs labeled as NTFS/UFS/QFS entities by MDC
- Not a File System
- Third-party product dependent

DataPlow: Basic Architecture

- Distributed metadata
  - Higher level information managed by metadata server
  - Extent-level data stored directly on the shared storage
  - SAN Clients allocate/deallocate their own metadata and real data
- Metadata server
  - Standalone host or combined with SAN client
  - IRIX or Sun platform
  - Metadata stored on host's local file system disk
- Split data versus control flow
  - Data files and extent-level data moved over fibre channel links
  - Control passed between SAN clients and FSS via LAN
- Third party volume management
- Normal file system utilities supported

GFS: Basic Architecture

- Distributed metadata
  - Metadata data stored on the shared storage
- Disk based file locks (dlocks)
  - Shared locks (read) prohibit writes
    - Allows concurrent reads
  - Exclusive locks (write) prohibits concurrent write/read
  - Aggressive caching schemes for both locks and data
  - LAN connectivity between SAN clients for lock release
- GFS specific volume management (pool)

File Systems Under Evaluation

- MountainGate Imaging Systems Corp., Inc.
  - CentraVision™ File System (CVFS)
  - Acquired by ADIC
- Mercury Computer Systems, Inc.
  - SANergy™
  - Unit acquired by Tivoli Systems
- DataPlow, Inc.
  - DataPlow™ SAN File System (SFS)
- Global File System (GFS)
  - University of Minnesota with NASA/DoD funding

Future Efforts

- Operational stress testing
  - Fabric routing
  - Failover
- Backup evaluation
- File system, archive and HSM integration

Research Testbed Expansion

- Computers
  - SGI Origin2000
  - DEC
  - Sun
  - IBM
  - Dell
- Operating Systems
  - Solaris
  - Tru64
  - AIX
- Interconnects
  - FC-to-SCSI bridge
- Mass Storage
  - StorageTek
  - Exabyte
  - HP SureStore
  - Qualstar
  - IBM
- RAID
  - Sun
- Software
  - Back-up
  - HSM

Test Programs To Be Used

- Postmark
  - Designed for small network traffic (NetApps)
  - Anomalous test data; under study
- DataPlow special test
  - Origin binary only; Solaris not yet obtained
  - Synchronization feature of interest to measure back-off/retry scenarios
- VXbench
  - Solaris and NT only

Challenges in Testing

- Changing hardware baseline:
  - Stable environment needed to compare SAN file systems
    - Component performance varies
    - Components don't all interoperate
    - Failover causes disappearing performance without warnings… manual action required to restore
  - Restoring baseline configurations after an HBA change, for example, can be very time-consuming
- Unexpected 3rd party software interactions
  - Made SANergy testing difficult

CVFS: Basic Features Tested

- 64-bit file system - files larger than 2GB supported
- Data stored in CVFS file format: supports big/little endian conversion concerns between NT & Origins
- NFS export capabilities (tested separately; NFS V3 only)
- Centralized metadata management: File System Services
  - Tested on NT and IRIX platform; performance similar for 3 clients; vendors models say to 70 clients possible
  - Metadata stored on host's local file system only
  - Extent-based file system, aggressive allocation supported
- Completely Split data versus control flow
  - Control data passed between SAN clients and FSS host via LAN
  - Data files moved over fibre channel links
- Volume management: self-contained, part of cvfs

Cvfs Blob Efficiency- Small LUNs

4GB/4MB io (Prisa/Ancor/ write cache on)

Stream 1	Stream 2	Aggregate	Same File	FileSys
53.97	38.66	92.63	r+w (1-O2K)	2 x R3 4+1
53.62	36.97	90.59	r+w (1-O2K)	1 SP
28.79	28.8	57.59	r+r (2-O2Ks)	BS=4kB
24.36	24.31	48.67	r+w (2-O2Ks)	StBr=64KB
42.52			r (NTFSS)	StBr=16Kb

Cvfs - Scaling (Small LUNs)
(Prisa/Ancor/write cache off)

Stream 1	I/O Blk Size	#LUNs/#SPs	Direction	NTFSS
42.3	1MB	2+2	read	stBr=16KB
5.62	1MB	2+2	write	BS=4K
23.09	1MB	1+1	read	
2.47	1MB	1+1	write	

Cvfs - I/O Block Size Effect
(Prisa/Ancor/write cache on)

I/O Block	# Blocks Mvd	Direction	Rate (MB/s)	NTFSS
512 KB	2000	write	17.25	stBr=16KB
512 KB	2	write	10.98	BS=4K
512 KB	10	write	17.18	1 - R3 4+1
512 KB	20	write	17.84	
512 KB	2000	read	23.48	
1 MB	1000	write	21.24	
1 MB	1000	read	25.22	
2 MB	500	write	21.92	
2 MB	500	read	25.68	
4 MB	250	write	22.47	
4 MB	250	read	30.15	
8 MB	250	read	29.71	

CVFS: Features and Releases

- Strengths Tested
  - Aimed at direct I/O, large files, large block I/Os
  - File system block size, amount of data to each LUN studied
  - True multi-LUN striping tested
  - Efficiency and performance compare well to XFS on Origins
- Further testing:
  - Additional clients: Linux and Solaris
  - Memory-mapped I/O (better small file performance)
  - Concatenated as well as striped file system testing
  - Combination with HSM products; potential for serverless transfers

DataPlow - Striped vs Concatenated
(lmdd - 4 x RAID-3 8+1 LUNs - 128x512)

FS Type	Menagea	Menageb	Aggregate
Striped	66.04 - wr	144.33 - rd	-
Concatenated	76.91 - wr	87.72 - rd	-
Striped	69.90 - rd	69.80 - rd	139.70 MB/s
Concatenated	38.08 - rd	38.07 - rd	76.15 MB/s
Striped	33.34 - wr	33.28 - wr	66.62 MB/s
Concatenated	23.62 - rd	21.46 - wr	45.08 MB/s